Talk with Christoph Lampert

Variational Deep Learning

Ashuha Arseniy[1,2], Dmitriy Molchanov[1,3]

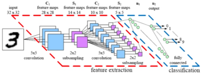
Bayesian Research Group[1], MIPT[2], SkolTech[3]

p(B|A)yesgroup.ru

ars.ashuha@gmail.com

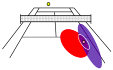
September 9, 2016

Two Stream of Machine Learning



Deep Learning

- \+ Rich non-linear models for classification and sequence prediction.
- \+ Scalable learning using stochastic approximations and conceptually simple.
- \+ Easily composable with other gradient-based methods
- − Only point estimates
- − Hard to score models, do model selection and complexity penalisation.



Bayesian Reasoning

- − Mainly conjugate and linear models
- − Potentially intractable inference leading to expensive computation or long simulation times.
- \+ Unified framework for model building, inference, prediction and decision making
- \+ Explicit accounting for uncertainty and variability of outcomes
- \+ Robust to overfitting; tools for model selection and composition.

Review and Limitations of Deep Learning

- We all know well the linear models:

$$\nu = \mathbf{w}^t x + b, \quad p(y|x) = p(y|g(\nu); \theta)$$

  - The basic function can be any linear function, e.g., affine, convolution
  - $g(\cdot)$ is a function that we'll refer to as an activation function
- Recursive composition generalized linear functions give a Deep NN

$$NuralNet(x) = g_n(W_n \cdot \ldots \cdot g_1(W_1 \cdot x))$$

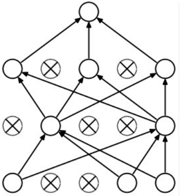
- While training we usually optimize Maximum Likelihood Estimation

A general framework for building non-linear, parametric models

Problem: Overfitting of MLE leading to limited generalisation

Regularization of Deep Neural Nets

- Regularization is essential to overcome overfitting
- A wide range of available regularization techniques:
    - Large data sets
    - Input noise and data augmentation
    - L2 / L1 regularization
    - Binary or Gaussian Dropout
    - Batch normalization



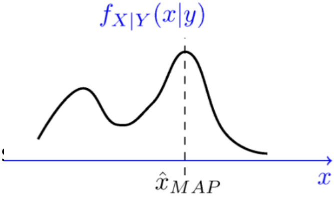
More robust loss function using MAP estimation instead

\+ Power of MAP estimators is that they provide some robustness to overfitting

\+ Automatic determination of feature relevance

\+ Can generate of frequentest confidence intervals

− Creation of sensitivities to parametrization



Review and Limitations of Bayesian Reasoning

Our goal is to learn a model with weights w of the $p(y|x, w)$

Let's use Bayesian toolbox:

1. We want to give posterior distribution $p(w|D) = p(w)p(D|w)/p(D)$
2. But Bayes rule involves computationally intractable integrals
3. Therefore we will introduce parametric family $q_\phi(w)$
4. Let's tune parameters such that $q_\phi(w)$ will be a close to $p(w|D)$
5. Usually by maximize Variational Lower Bound

$$\mathcal{L}(\phi) = \sum_{(x,y)\in D} \mathbb{E}_{q_\phi} log\ p(y|x, w) - D_{KL}(q_\phi(w), p_{prior}(w))$$

likelihood expectation regularizer

But, there is problem to apply this approach for Deep Nets

- $p(y|x, w)$ is Neural Net, so $\mathbb{E}_{q_\phi} log\ p(y|x, w)$ became intractable
- $\sum_{(x,y)\in D} \cdot$ contain sum over data, so EM-like algo cannot be applied

Approximate Bayesian Inference

The idea of stochastic optimization is quite simple, we will use:

- gradient based methods
- unbiased gradient estimation instead of true gradient

Can we apply this approach to optimize variational lower bound?

$$\mathcal{L}(\phi) = \sum_{(x,y)\in D} \mathbb{E}_{q_\phi} log\ p(y|x, w) - D_{KL}(q_\phi(w), p_{prior}(w))$$

1. We can't take gradient by parameters by naive way.

$$\nabla_\phi \mathbb{E}_{q_\phi} log\ p(y|x, w) \neq \mathbb{E}_{q_\phi} \nabla_\phi log\ p(y|x, w)$$

2. Let's use re-parametrization

$$\nabla_\phi \mathbb{E}_{q_\phi} log\ p(y|x, w) = \mathbb{E}_{N(\epsilon|0,1)} \nabla_\phi log\ p(y|x, w = f(\epsilon, \phi))$$

3. For example $q_\phi = N(\phi_1, \phi_2^2)$ then $f(\epsilon, \phi) = \phi_1 + \phi_2 \cdot \epsilon$
4. We can compute estimation of grad using double stochastic inference

Variational Dropout

- Affine layer with parameters matrix W looks like

$$B = A \cdot W$$

- Dropout is Bernoulli noise on input matrix and scaling

$$B = (A \odot (\xi/(1-p)))W \quad \xi \sim Bernoulli(p)$$

- Gaussian Noise with the same mean and variance works as well

$$B = (A \odot \xi)W \quad \xi \sim Gaussian(1, p/(1-p)) = Gaussian(1, \alpha)$$

- It correspond to normal posterior distribution over weights

$$N(\mu, \sigma^2) = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim Gaussian(1, \alpha)$$

$$B_{ij} = (A_i \odot \xi)W^j = \sum_t A_{it} \cdot (1 + \sqrt{\alpha} \cdot \epsilon) \cdot W_{tj} =$$

$$= \sum_t N(A_{it}, \alpha A_{it}^2) \cdot W_{tj} = \sum_t A_{it} \cdot N(W_{tj}, \alpha W_{tj}^2)$$

$$q_{\alpha,\phi}(w_{ij}) = N(\phi_{ij}, \alpha \phi_{ij}^2)$$

Variational Dropout

During training neural net with dropout we optimize wrt ϕ with fixed α

$$\sum_{(x,y)\in D} \mathbb{E}_{q_{\alpha,\phi}} log\ p(y|x,w) \to \max_{\phi}, \quad q_{\alpha,\phi}(w_{ij}) = N(\phi_{ij}, \alpha\phi_{ij}^2)$$

With Log-uniform prior on w_{ij}

$$p(log\ |w_{ij}|) \propto c$$

divergence does not depend on ϕ

$$-D_{KL}(q_{\alpha,\phi}(w_{ij})||p(w_{ij})) = const + 0.5 \cdot log(a) + E_{\epsilon\sim N(1,\alpha)} log|\epsilon|$$

Thus during dropout training we optimize variational lower bound w.r.t. ϕ

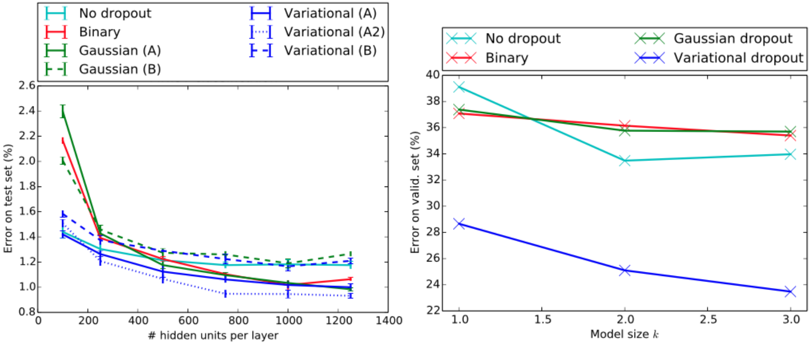
$$\sum_{(x,y)\in D} \mathbb{E}_{q_{\alpha,\phi}} log\ p(y|x,w) - D_{KL}(q_{\alpha,\phi}(w_{ij})||p(w_{ij})) \to \max_{\phi}$$

Dropout is the special case of Bayesian Regularization

This is important because:

1. we can train personal alpha for weight, features or layer
2. physical interpretation – number of significant digits

Experiments Result



(a) Classification error on the MNIST dataset (b) Classification error on the CIFAR-10 dataset

The experimental part of the paper was quite strange:

1. Alphas are clipped at 1: $\alpha < 1, \quad p < 0.5$
2. Training method for alphas and alpha sharing scheme are not specified
3. The KL divergence was divided by 3 to prevent underfitting

Our: ARD with VDO in Linear Models [D. Molchanov]

- Relevance Vector Machine

$$p(t|x,w) = \sigma(tw^t x)$$

$$p(w|a) = N(w|0, diag(\alpha_1^{-1}, \dots, \alpha_n^{-1}))$$

and determine α_i by optimizing evidence maximization

$$P(X|\alpha) = \int p(t|x,w)p(w|\alpha)dw \to \max_{\alpha}$$

if feature j is irrelevant $\alpha_j \to +\infty, w_j \to 0$.

- Relevance Determination with Variational Drop Out

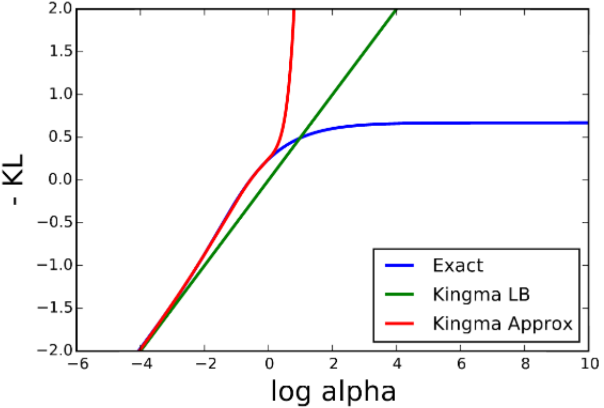
$$\sum_{(x,y)\in D} \mathbb{E}_{q_{\alpha,\phi}} log\ \sigma(yw^t x) - D_{KL}(q_{\alpha,\phi}(w_{ij})||p(w_{ij})) \to \max_{\phi}$$

While we optimize Variational Lower Bound with log-Uniform Prior

  - automatic relevance determination still exists
  - tuning parameters of the prior distribution isn't necessary!!!

Our: Divergence Estimation in VDO [D. Molchanov]

Approx. of the D_{KL} which was obtained in the article is only true if $\alpha < 1$



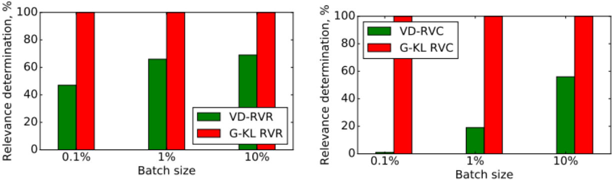
It was obtained by using sampling method.

$$-D_{KL}(q_{\alpha,\phi}(w_{ij})||p(w_{ij})) = const + 0.5 \cdot log(a) + E_{\epsilon\sim N(1,\alpha)} log|\epsilon|$$

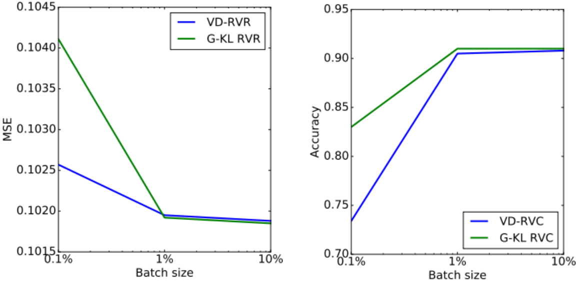
Our: ARD with VDO in Linear Models [D. Molchanov]

Synthetic data, 100000 objects, 10 relivant, 90 irrelivant features.

- Reliance determination (Regression, Classification)



- Error Regression, Accuracy Classification



Our: More Features [D. Molchanov]

1000 objects, 10 relevant features, 990 irrelevant features.

Linear regression:

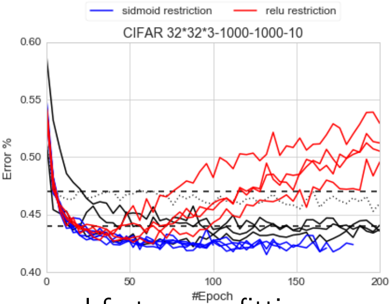
Method	Test MSE	Relevance determination
VD-RVR	0.119	816/990
G-KL RVR	0.128	990/990

Logistic regression:

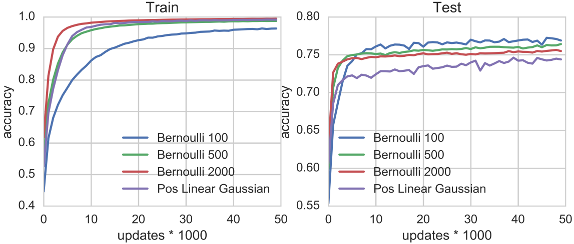
Method	Test accuracy	Relevance determination
VD-RVC	0.916	984/990
G-KL RVC	0.907	983/990

Our: VDO picture classification without clipping

VDO picture classification without clipping



variance of gradients coursed faster over-fitting



Our: future work

- **Automatic Reliance determination**

  Determine personal alpha per layer/features with ARD
- **Incremental Learning**

  Using posterior distribution as a prior for next portion of data
- **Another regularization scheme**

  To use Log-Normal distribution to introduce unsymmetrical noise.

References

- Diederik P. Kingma, Tim Salimans, Max Welling: Variational Dropout and the Local Reparameterization Trick, arxiv.org/abs/1506.02557
- Diederik P Kingma, Max Welling: Auto-Encoding Variational Bayes, arxiv.org/abs/1312.6114
- Shakir Mohamed, http://shakirm.com/, http://blog.shakirm.com/wp-content/uploads/2015/10/Bayes˙Deep.pdf